APPENDIX E

U.S. Department of State
Bureau of Consular Affairs

Consular Lookout and Support System—Enhanced (CLASS-E)

Presentation to
CA / EX / CSD

Advanced Name Classifier for CLASS-E (ANC-E)

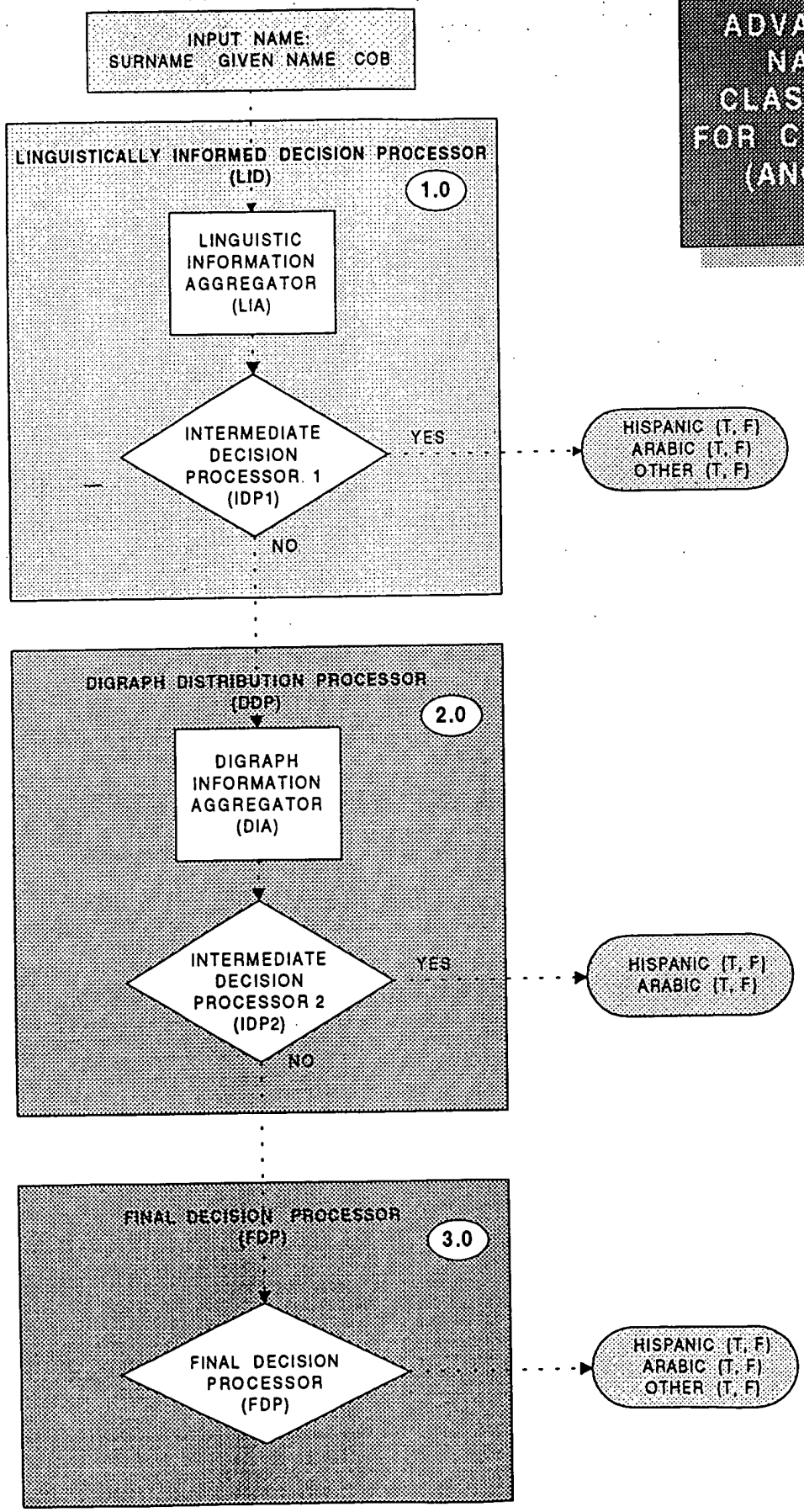
and

Hispanic Name Search Algorithm for CLASS-E (HNA-E)

September 18, 1997

Language Analysis Systems, Inc.
2214 Rock Hill Road—Herndon, VA—20170

ADVANCED NAME CLASSIFIER FOR CLASS - E (ANC - E)

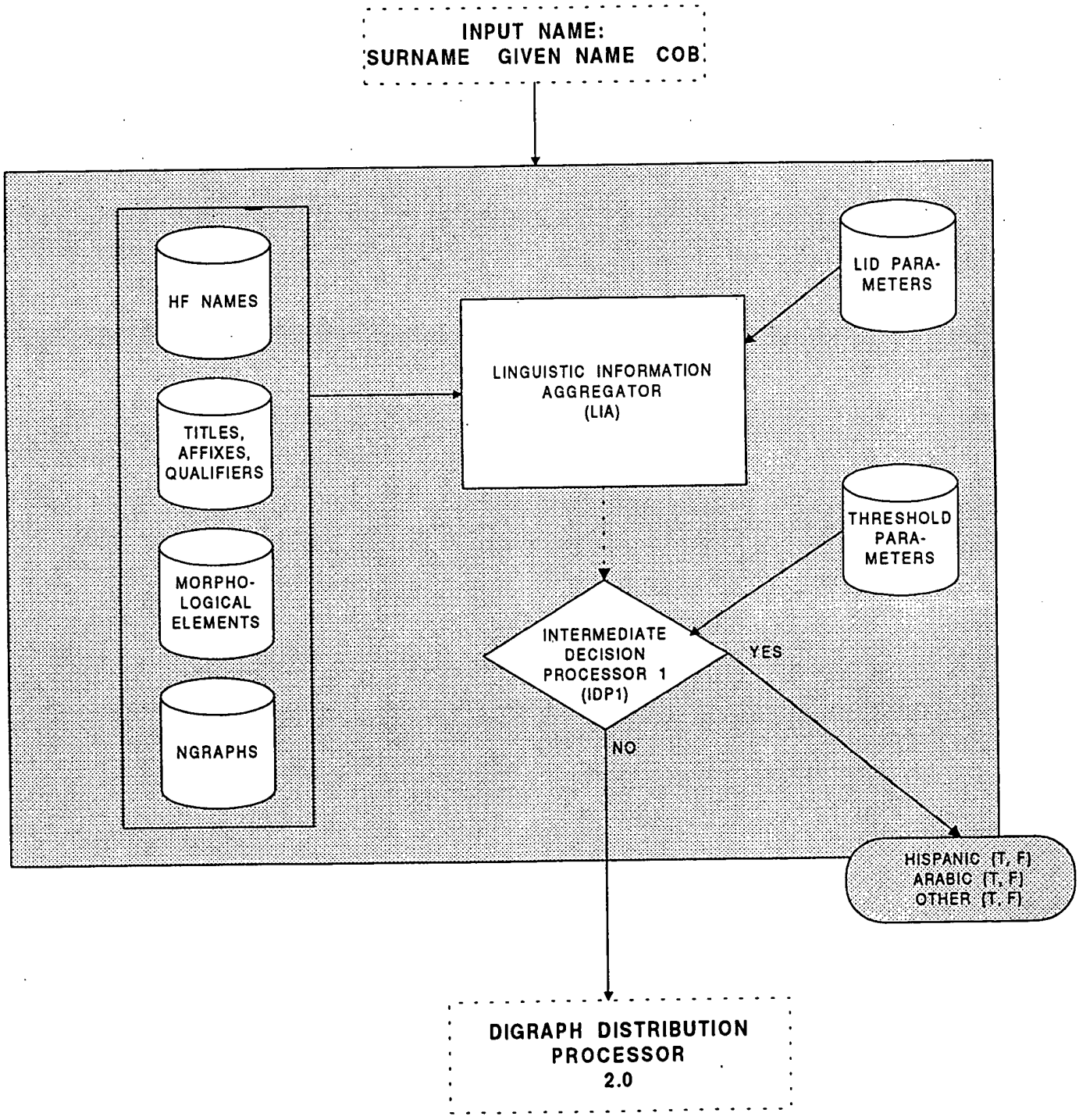


Language Analysis Systems, Inc.
hmb 9/16/97
CA/EX/CSD Presentation

LAS

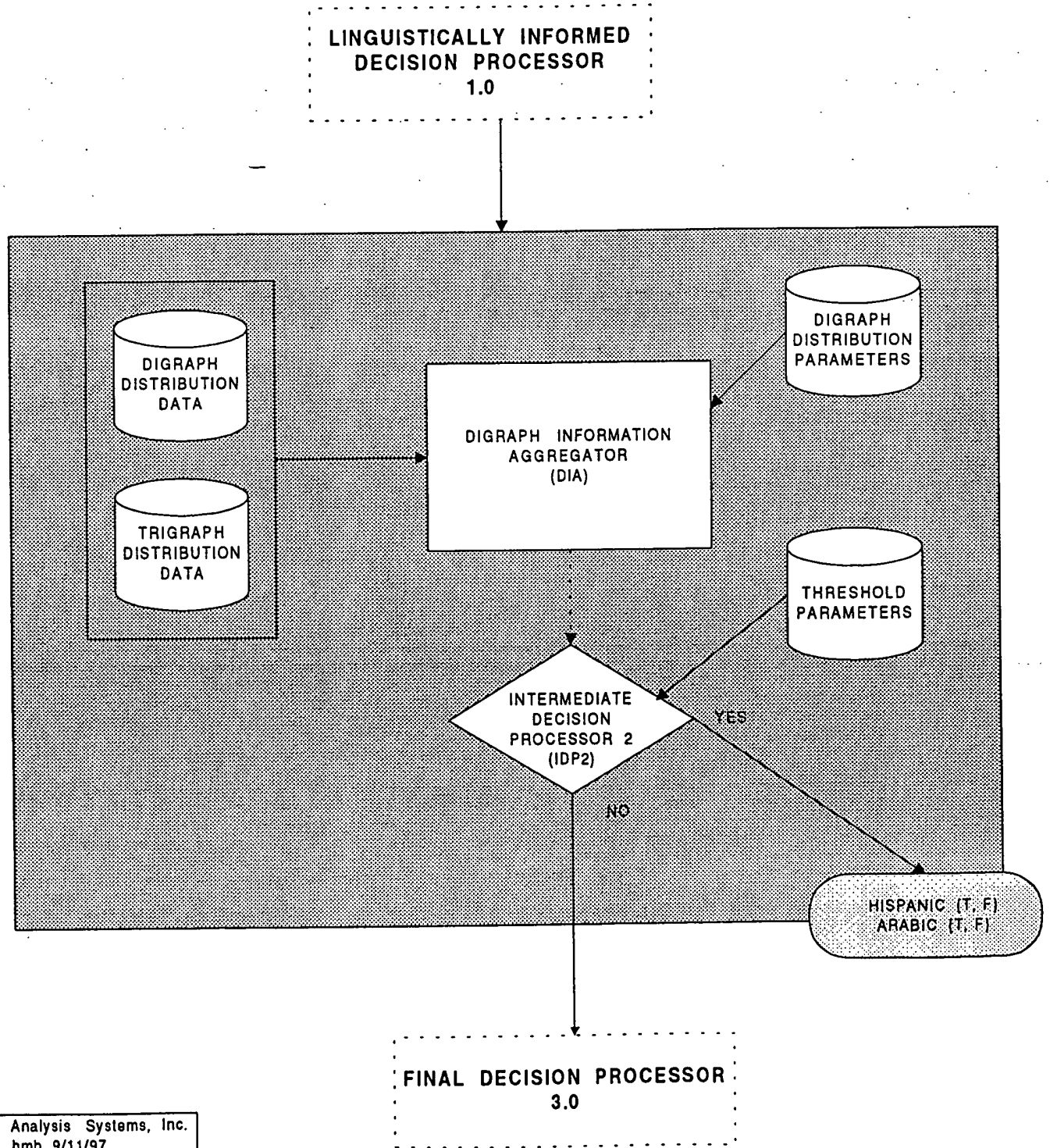
1.0

LINGUISTICALLY INFORMED DECISION PROCESSOR (LID)



2.0

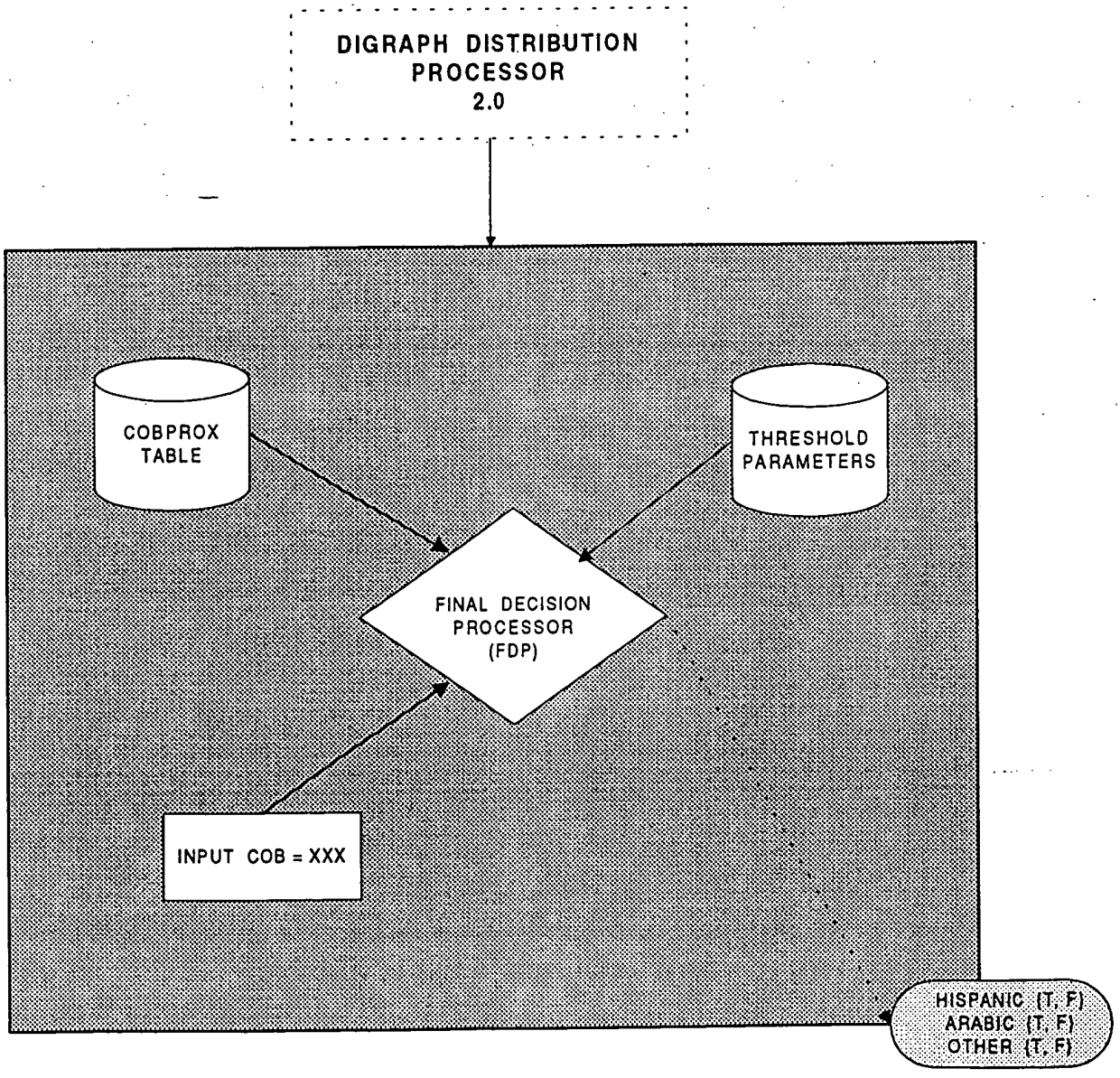
DIGRAPH DISTRIBUTION PROCESSOR (DDP)



Language Analysis Systems, Inc.
hmb 9/11/97
CA/EX/CSD Presentation

LAS

FINAL DECISION PROCESSOR (FDP)



ADVANCED NAME CLASSIFIER (ANC - E) EXAMPLE

LINGUISTICALLY INFORMED DECISION PROCESSOR (LID)

Factors:	In Field: 10 OutField: 8	In Field: 8 OutField: 6	In FieldSN: 5 OutFieldSN: 3 In FieldGN: 4 OutFieldGN: 2	In FieldSN: 3 OutFieldSN: 2 In FieldGN: 2 OutFieldGN: 1	In FieldSN: 5 OutFieldSN: 3 In FieldGN: 4 OutFieldGN: 2

High Frequency SN			
H	S	Garcia	3
H	S	Salazar	2
H	S	Sambrano	1
O	S	Greco	5
O	B	Giuliano	2
O	S	Silvestri	1

High Frequency GN			
H	G	Jose	3
H	G	Francisco	2
H	G	Mario	1
O	B	Luigi	3
O	G	Antonio	2
O	G	Adalberto	1

Prefixes			
H	S	de	1
H	S	la	1
H	S	las	1
O	B	il	1
O	B	el	1
O	B	lo	1

N-Grams			
H	S	-ndez	3
H	S	-guez	2
H	S	-illo	1
O	S	-illo	3
O	B	-ini	2
O	S	-agio	1

Morphology			
A	S	adin	3
A	S	eddin	2
A	S	uddin	1
O	S	etto	5
O	S	etti	4
O	S	ini	2

N.B.:The data shown here are for the purpose of illustration only and do not necessarily reflect actual table values.

DELGADILLO DE GARCIA, JOSE ANTONIO

	-ILLO	ANTONIO	DE	GARCIA	JOSE		
Hispanic:	(3*1)		(5*1)	(10*3)	(8*3)	=	62
Arabic:	--	--	--	--	--	=	0
Other:	(3*3)	(8*2)	--	--	--	=	25

LID_Threshold: Hispanic 65; Arabic 57; Other 56
Hispanic: F; Arabic: F; Other: F
(N.B. Values for illustration only.)

DIGRAPH DISTRIBUTION PROCESSOR (DDP)

DELGADILLO DE GARCIA, JOSE ANTONIO

DIGRAPHS		
A	EZ	-1.0422
A	NT	22.8733
A	BD	38.7221
H	BD	1.0572
H	EZ	42.5947
H	RI	16.1242

TRIGRAPHS		
A	EZ#	-10.0422
A	#NT	-48.1743
A	BD#	48.4551
H	BD#	-32.1742
H	EZ#	47.5327
H	#RI	11.1242

Digraph/Trigraph Scores:

HISPANIC: 44.2331
ARABIC: -32.8765

DI_Threshold: Hispanic 45.1116; Arabic 1.4532
Hispanic: F; Arabic: F

FINAL DECISION PROCESSOR (FDP)

DELGADILLO DE GARCIA, JOSE ANTONIO

COB = COL

COB	PART	COB2
COL	H	COL
COL	H	VENE
COL	H	BOL
EGYP	A	EGYP
EGYP	A	UAE

UNDER_DI_THRESHOLD: Hispanic 5; Arabic 10
UNDER_LID_THRESHOLD: Hispanic 6; Arabic 8; Other 3

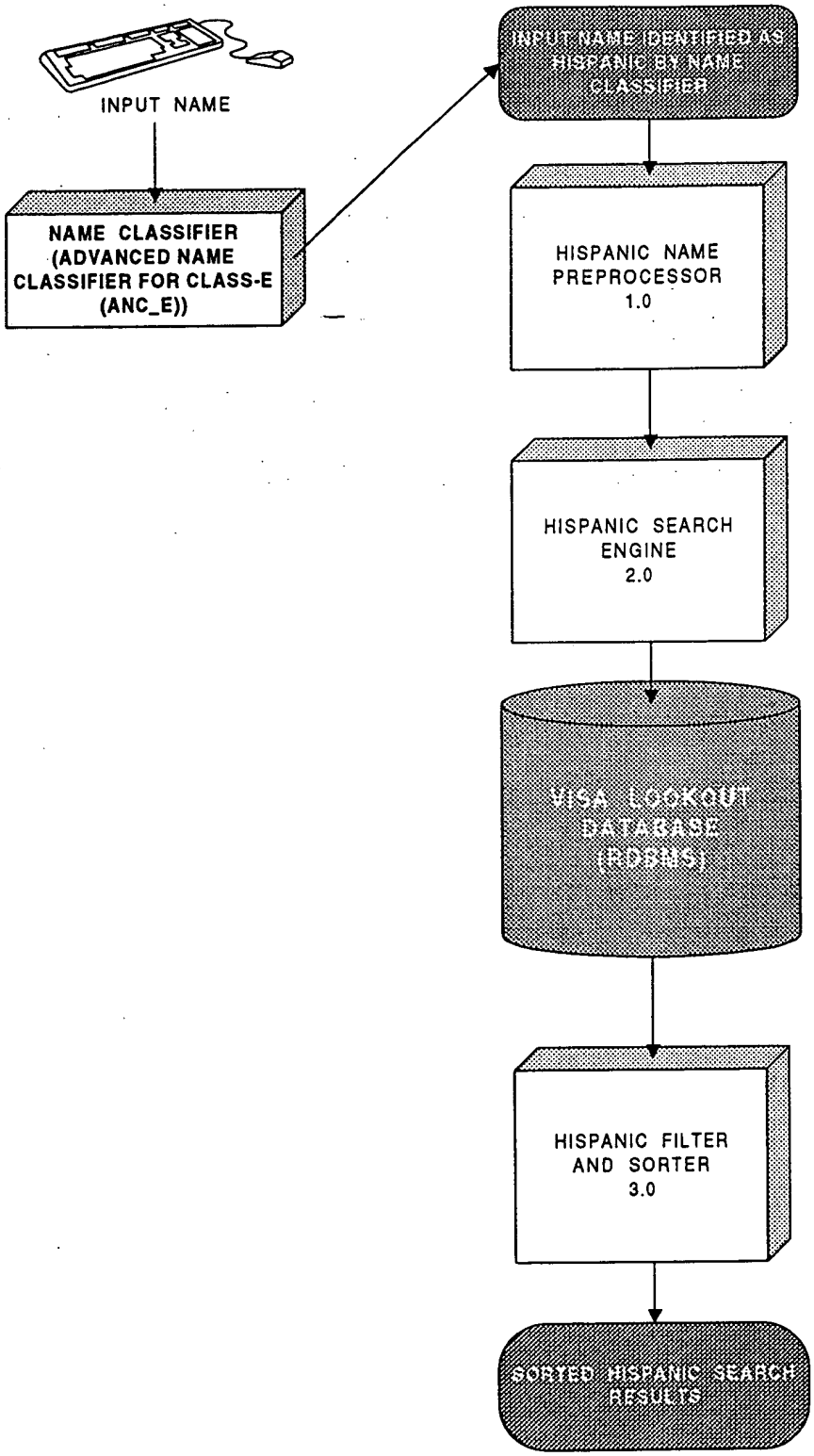
HISPANIC → YES
UNDER_DI_THRESHOLD – yes
UNDER_LID_THRESHOLD – yes
COB - H

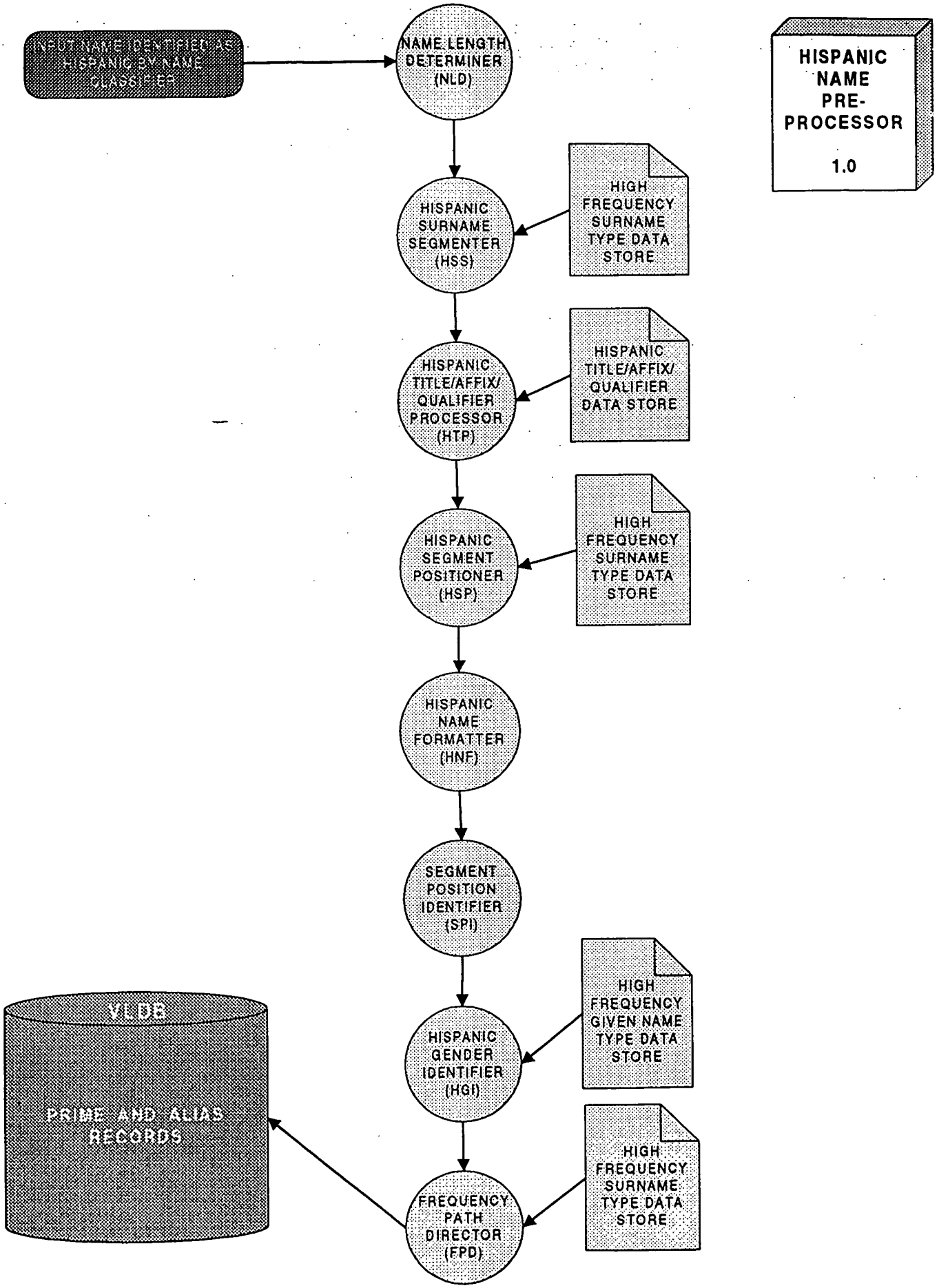
ARABIC → NO
UNDER_DI_THRESHOLD – no
UNDER_LID_THRESHOLD – no
COB - H

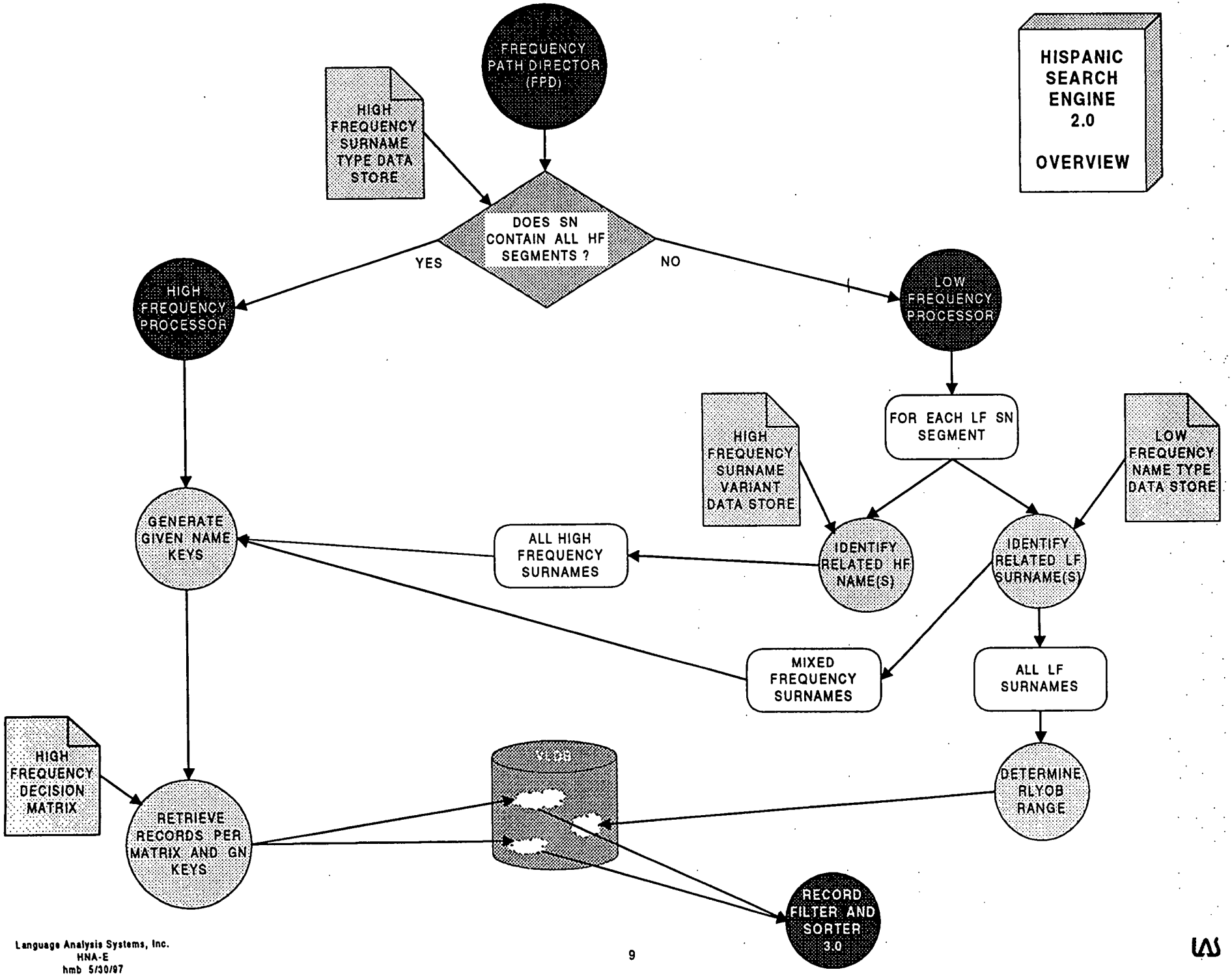
OTHER → NO
UNDER_LID_THRESHOLD – no
COB - H

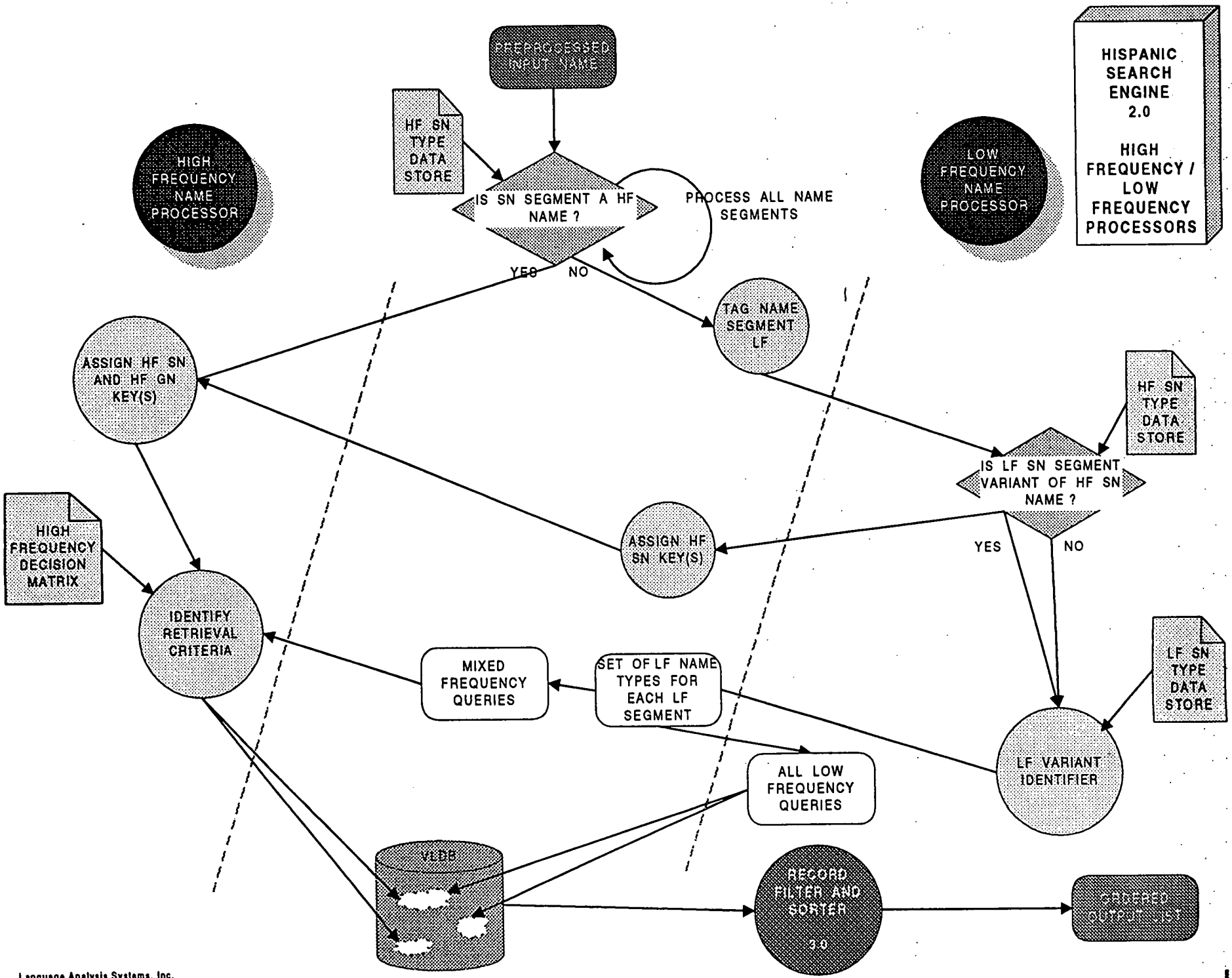
ADVANCED HISPANIC NAME SEARCH ALGORITHM for CLASS - E (HNA-E)

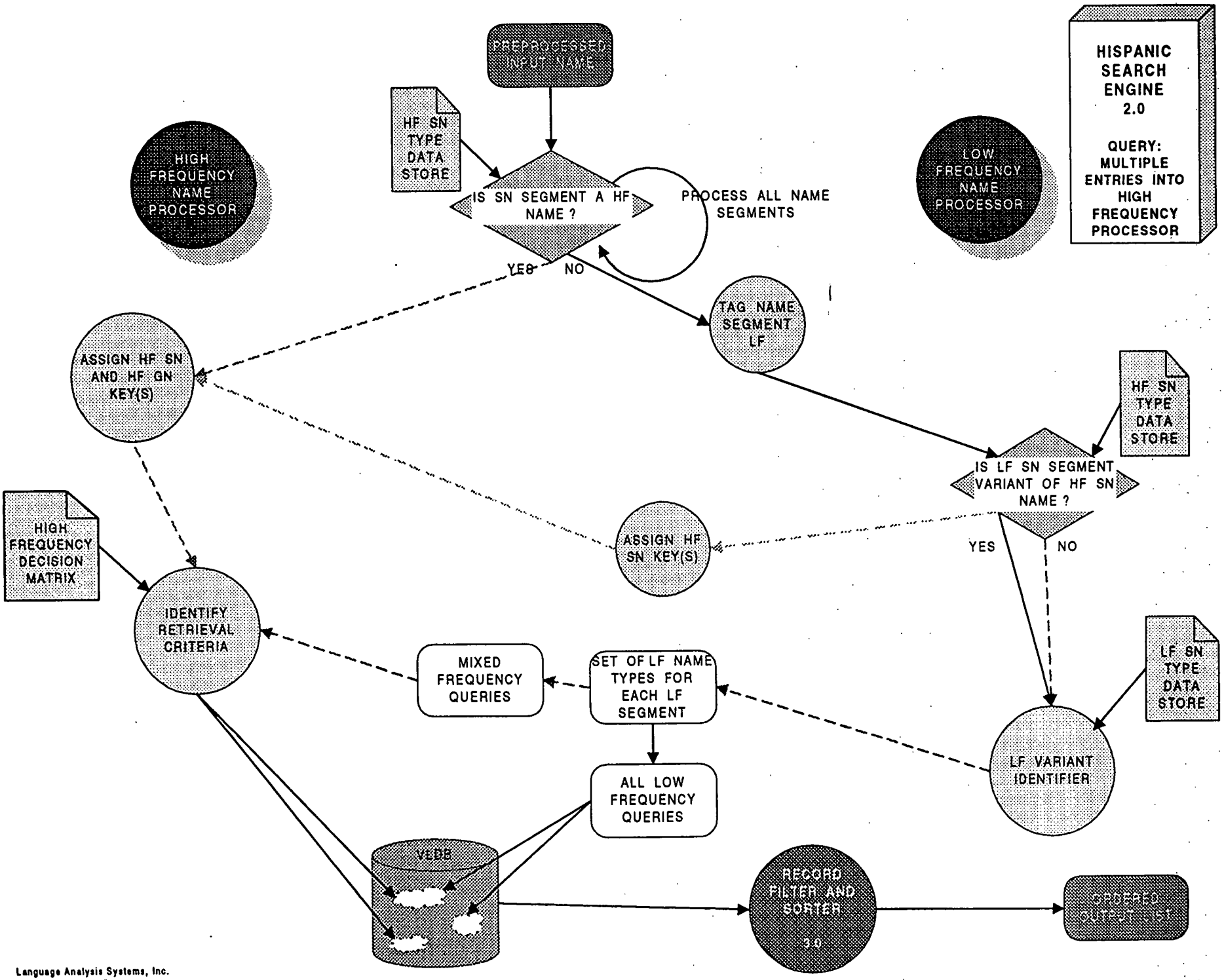
PROCESS FLOW



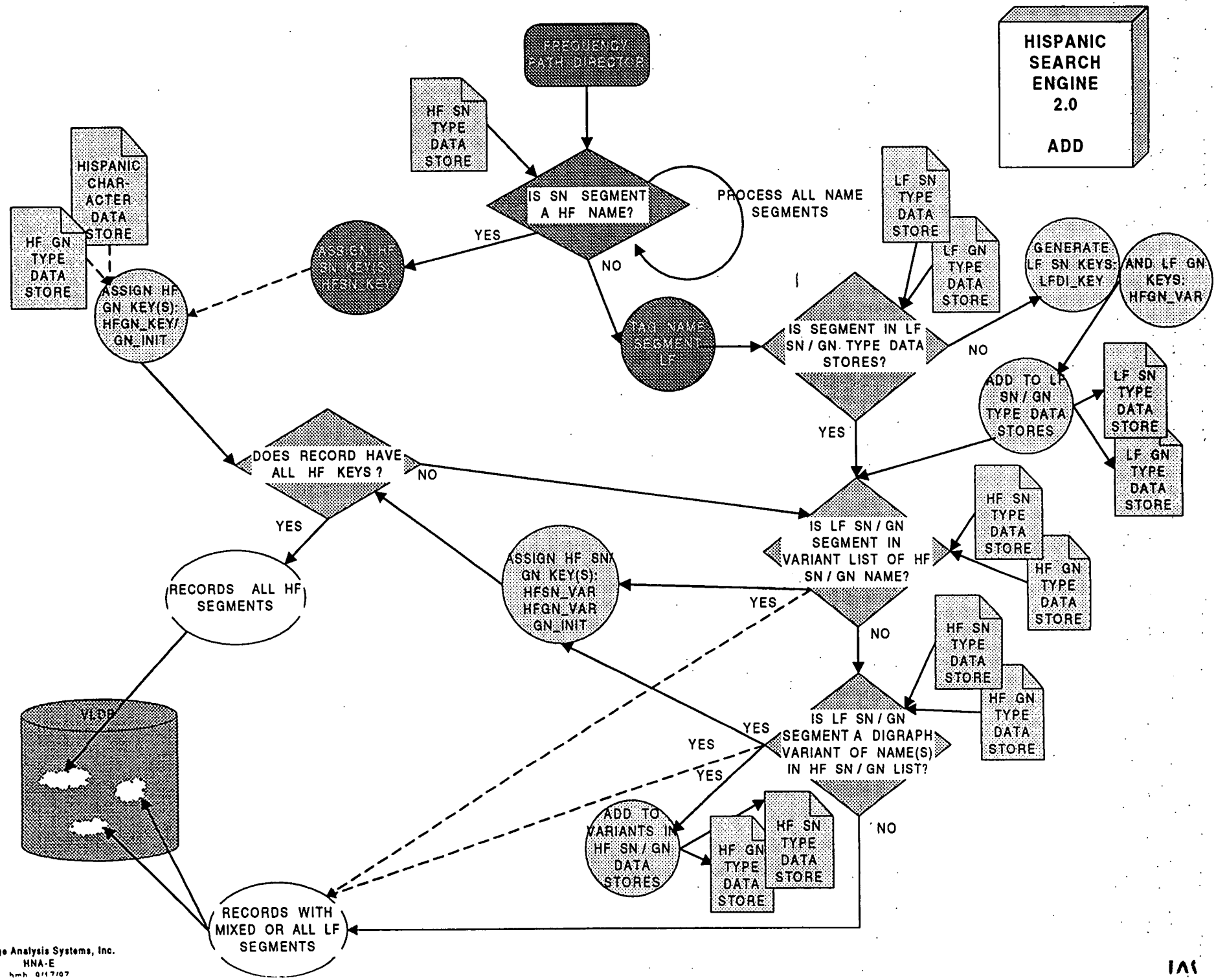

INPUT NAME IDENTIFIED AS HISPANIC BY NAME CLASSIFIER
NAME LENGTH DETERMINER (NLD)
HISPANIC NAME PRE-PROCESSOR 1.0
HISPANIC SURNAME SEGMENTER (HSS)
HIGH FREQUENCY SURNAME TYPE DATA STORE
HISPANIC TITLE/AFFIX/ QUALIFIER PROCESSOR (HTP)
HISPANIC TITLE/AFFIX/ QUALIFIER DATA STORE
HISPANIC SEGMENT POSITIONER (HSP)
HIGH FREQUENCY SURNAME TYPE DATA STORE
HISPANIC NAME FORMATTER (HNF)
SEGMENT POSITION IDENTIFIER (SPI)
HISPANIC GENDER IDENTIFIER (HGI)
HIGH FREQUENCY GIVEN NAME TYPE DATA STORE
FREQUENCY PATH DIRECTOR (FPD)
HIGH FREQUENCY SURNAME TYPE DATA STORE
VLDB
PRIME AND ALIAS RECORDS


HISPANIC SEARCH ENGINE 2.0
OVERVIEW
FREQUENCY PATH DIRECTOR (FPD)
HIGH FREQUENCY SURNAME TYPE DATA STORE
DOES SN CONTAIN ALL HF SEGMENTS ?
YES
NO
HIGH FREQUENCY PROCESSOR
LOW FREQUENCY PROCESSOR
FOR EACH LF SN SEGMENT
HIGH FREQUENCY SURNAME VARIANT DATA STORE
LOW FREQUENCY NAME TYPE DATA STORE
GENERATE GIVEN NAME KEYS
ALL HIGH FREQUENCY SURNAMES
IDENTIFY RELATED HF NAME(S)
IDENTIFY RELATED LF SURNAME(S)
MIXED FREQUENCY SURNAMES
ALL LF SURNAMES
HIGH FREQUENCY DECISION MATRIX
RETRIEVE RECORDS PER MATRIX AND GN KEYS
RLYOB
DETERMINE RLYOB RANGE
RECORD FILTER AND SORTER 3.0

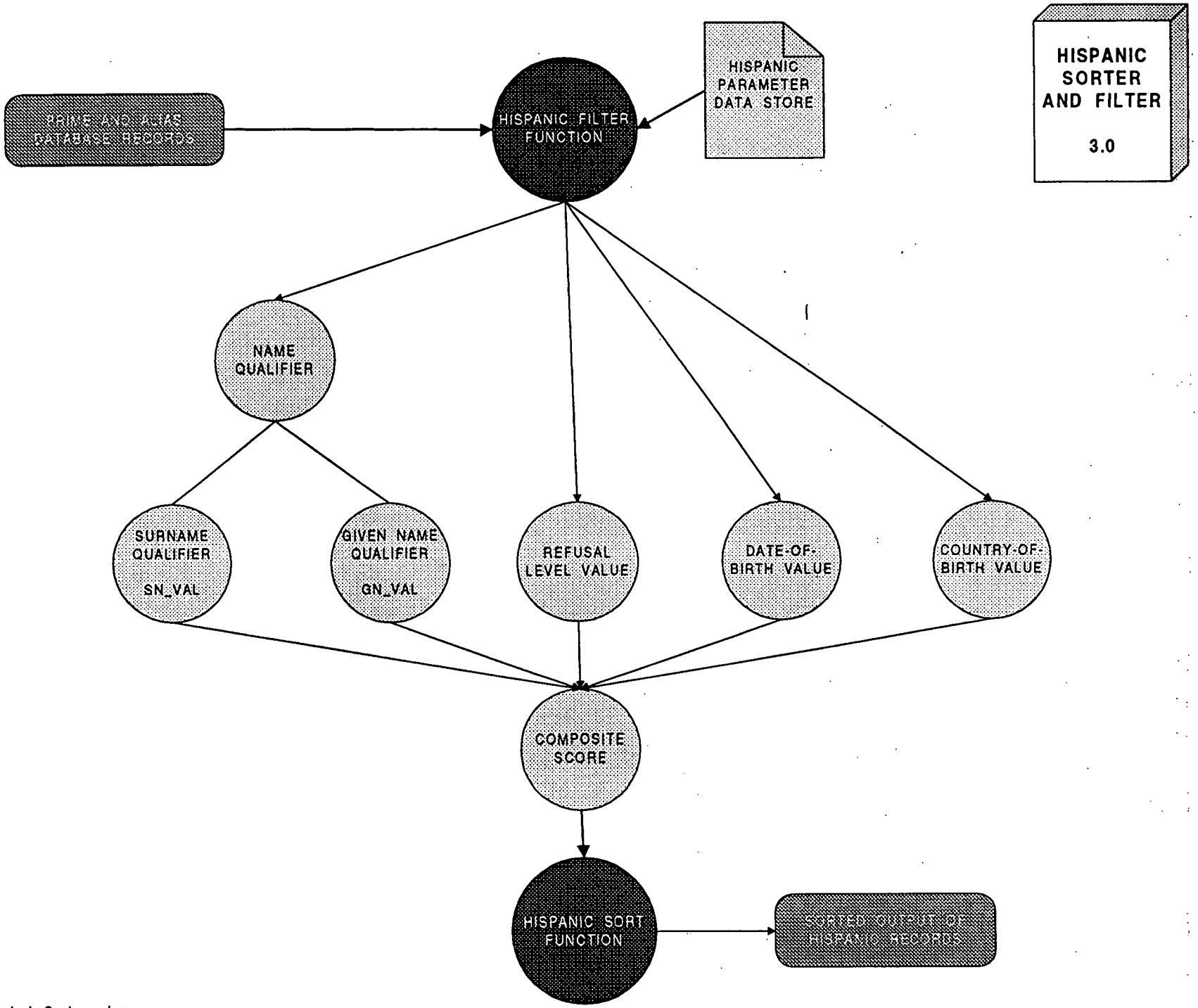

HISPANIC SEARCH ENGINE 2.0
HIGH FREQUENCY / LOW FREQUENCY PROCESSORS
PREPROCESSED INPUT NAME
HF SN TYPE DATA STORE
HIGH FREQUENCY NAME PROCESSOR
LOW FREQUENCY NAME PROCESSOR
IS SN SEGMENT A HF NAME ?
PROCESS ALL NAME SEGMENTS
YES
NO
TAG NAME SEGMENT LF
ASSIGN HF SN AND HF GN KEY(S)
HF SN TYPE DATA STORE
IS LF SN SEGMENT VARIANT OF HF SN NAME ?
HIGH FREQUENCY DECISION MATRIX
ASSIGN HF SN KEY(S)
YES
NO
IDENTIFY RETRIEVAL CRITERIA
MIXED FREQUENCY QUERIES
SET OF LF NAME TYPES FOR EACH LF SEGMENT
LF SN TYPE DATA STORE
LF VARIANT IDENTIFIER
ALL LOW FREQUENCY QUERIES
VLDB
RECORD FILTER AND SORTER 3.0
CRIBBED OUTPUT LIST



Language Analysis Systems, Inc.
HNA-E
hmb 9/15/97



Language Analysis Systems, Inc.
HNA-E
hmh 9/17/97


HISPANIC SORTER AND FILTER
3.0
PRINCE AND ALIAS DATABASE RECORDS
HISPANIC FILTER FUNCTION
HISPANIC PARAMETER DATA STORE
NAME QUALIFIER
SURNAME QUALIFIER
SN_VAL
GIVEN NAME QUALIFIER
GN_VAL
REFUSAL LEVEL VALUE
DATE-OF-BIRTH VALUE
COUNTRY-OF-BIRTH VALUE
COMPOSITE SCORE
HISPANIC SORT FUNCTION
SORTED CURSUT OF HISPANIC RECORDS

HISPANIC NAMESEARCH ALGORITHM FOR CLASS-E (HNA - E)

FREQUENCY PATH DIRECTOR

- **FPD directs record based on frequency of SURNAME data only**
- ***ALL* surnames must be HFSN_TYPEs for record to go directly to HF Processor**
- **FDP assigns HFSN_KEY (SET_ID in HFST) to each high frequency surname**

High Frequency Surname Type (HFST) Data Store (Sample)

ID_NO	HFSN_TYPE	SET_ID
0001	GARCIA	0001
0002	RODRIGUEZ	0002
0003	HERNANDEZ	0003
0004	LOPEZ	0004
0005	MARTINEZ	0005
0006	GONZALEZ	0006
0007	PEREZ	0007
0008	SANCHEZ	0008
0009	RAMIREZ	0009
0010	GOMEZ	0010
0011	...	0011

GARCIA LOPEZ, ANTONIO JESUS
0001 0004

HIGH FREQUENCY PROCESSOR

BOMEZ PEREZ, JOSE WILLIAM
----- 0007

LOW FREQUENCY PROCESSOR

HIGH FREQUENCY PROCESSOR

- **High Frequency Processor assigns HFGN_KEY (SET_ID in HGT) to each High Frequency Given Name**

Hispanic Given Name Type (HGT) Data Store (Sample)

ID_NO	GN_TYPE	SET_ID	HI_FREQ	GNDR
0001	JOSE	0001	1	M
0002	MARIA	0002	1	F
0003	JUAN	0003	1	M
0004	LUIS	0004	1	M
0005	ANTONIO	0005	1	M
0006	CARLOS	0006	1	M
0007	JESUS	0007	1	M
0008	MANUEL	0008	1	M
0009	FRANCISCO	0009	1	M
0010	JORGE	0010	1	M
0011	...	0011		...
2367	DAGOBERTO	0000	0	M

GARCIA LOPEZ, **ANTONIO JESUS**

0001 0004 0005 0007

IF ALL NAME SEGMENTS HAVE BEEN ASSIGNED HFSN_KEYs AND HFGN_KEYs, THE HFP MATRIX ACCESSES THE HISPANIC DECISION MATRIX

- **HFP accesses Hispanic Decision Matrix for additional search criteria**

Hispanic Decision Matrix (HDM) (Sample)

	Single-Segment SN			Two-Segment SN							
QUERY SN FORMAT	A	A	A	AB	AB	AB	AB	AB	AB	AB	AB
DATABASE SN FORMATS	A	AB	BA	AB	BA	A	B	AC	CA	CB	BC
YR#	5	5	2	5	4	4	2	2	0	0	0
RL#	4	4	3	4	4	4	1	1	0	0	0
RGNDR	MFU	MFU	MFU	MFU	MFU	MFU	MFU	FU	MFU	MFU	MFU

GARCIA LOPEZ, ANTONIO JESUS

0001 0004 0005 0007

For example, records with following surnames and retrieval criteria would be retrieved:

NAME	YR#	RL#	RGNDR
GARCIA LOPEZ	5	4	MFU
LOPEZ GARCIA	4	4	MFU
GARCIA	4	4	MFU
LOPEZ	4	4	MFU
GARCIA MARTIN*	2	1	FU
MARTIN* GARCIA	2	1	MFU
MARTIN* LOPEZ	0	0	MFU
LOPEZ MARTIN*	0	0	MFU

* Any SN segment

SEND TO HISPANIC SEARCH ENGINE

All HFSN_KEYs

All HFGN_KEYS

All Search Criteria

*HISPANIC SEARCH ENGINE WILL RETRIEVE

- AN EXACT MATCH AND
- ALL RECORDS WITH SN KEYS WITH RETRIEVAL CRITERIA **AND** *AT LEAST* ONE HFGN_ KEY

LOW FREQUENCY PROCESSOR

(1) High Frequency Access

- **LFP determines if LF SN is variant of HF SN**
- **LFP assigns HFSN_VAR keys (ID_NO in HFSV) to SN that is variant of High Frequency Surname**

High Frequency Surname Variant (HFSV) Data Store (Sample)

ID_NO	HFSN_VAR	SET_ID	DI_VAL
032711	PEREZ	0007	1.00
032712	PERES	0007	0.67
032713	PEREZA	0007	0.77
016976	GOMEZ	0010	1.00
016977	GOMES	0010	0.67
016978	BOMEZ	0010	0.67

BOMEZ PEREZ, JOSE DAGOBERTO
016978 0007

LFP SENDS NAME WITH ALL HFSN_KEYs and HFSN_VAR KEYS TO *HFP*

- HFP WILL GENERATE GIVEN NAME KEYS
- HFP WILL IDENTIFY SEARCH CRITERIA IN HISPANIC DECISION MATRIX

(ADDITIONAL) GIVEN NAME KEYS

Hispanic Given Name Type (HGT) Data Store (Sample)

ID_NO	GN_TYPE	SET_ID	HI_FREQ	GNDR
0001	JOSE	0001	1	M
0002	MARIA	0002	1	F
0003	JUAN	0003	1	M
0004	LUIS	0004	1	M
0005	ANTONIO	0005	1	M
0006	CARLOS	0006	1	M
0007	JESUS	0007	1	M
0008	MANUEL	0008	1	M
0009	FRANCISCO	0009	1	M
0010	JORGE	0010	1	M
0011	...	0011		...
2367	DAGOBERTO	0000	0	M

EXAMPLE 1:

BOMEZ PEREZ, **JOSE DAGOBERTO**
016978 0007 0001 D

- JOSE WILL BE ASSIGNED HFGN_KEY
- DAGOBERTO (WHILE IN THE LIST OF GIVEN NAMES) IS A LOW FREQUENCY GN
- DAGOBERTO IS ASSIGNED A GN_INIT KEY OF D

EXAMPLE 2:

Hispanic Character (HCD) Data Store (Sample)

SET_ID	CHAR	CHAR_VAR
001	B	B
001	B	V
002	S	S
002	S	Z
004	C	C
004	C	S
...		
037	F	F
052	K	K
078	M	M
078	M	N
...		
093	J	J
093	J	H

BOMEZ PEREZ, **JOSSE DAGOBERTO**
016978 0007 ---- ----
(093) J/H (005) D

BOTH JOSSE AND DAGOBERTO ARE LOW FREQUENCY GIVEN NAMES

- BOTH JOSSE AND DAGOBERTO ARE ASSIGNED A GN_INIT KEY (093 (J/H) AND 005 (D))
- INITIAL VARIANTS ARE ACCESSED IN THE HISPANIC CHARACTER DATA STORE (HCD)

EXAMPLE OF RETRIEVAL WITH HIGH FREQUENCY SURNAME KEYS AND MIXED GN KEYS

Represents partial set of query patterns only (SLOPEZ will generate additional LF Keys)

QUERY #1	RODRIGUEZ	SLOPEZ	JOSSE	CARLOS	CRITERIA
HFSN_KEY	002				
HFSN_VAR Key		00976			
HFGN_KEY				0007	
GN_INIT Key(s)			041 (J, H)		
HDM FORMATS:					
1	RODRIGUEZ (002)	LOPEZ (000976)			YOB5, RL4, MFU, GN initial = J or H; or GN= 0007
2	LOPEZ	RODRIGUEZ			YOB4, RL4, MFU, GN initial = J or H; or GN= 0007
3	RODRIGUEZ				YOB4, RL4, MFU, GN initial = J or H; or GN= 0007
4	LOPEZ				YOB2, RL1, MFU, GN initial = J or H; or GN= 0007
5	RODRIGUEZ	*			YOB2, RL1, FU, GN initial = J or H; or GN= 0007
6	LOPEZ	*			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
7	*	RODRIGUEZ			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
8	*	LOPEZ			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007

LOW FREQUENCY PROCESSOR

(2) Low Frequency Surname Processor

(Includes all SN *not* identified as HF, even HF Variant SN)

Low Frequency Surname Type (LFST) Data Store (Sample)

ID_NO	LFSN_TYPE	LFDIKEY
000001	AALVAREZ	AA1
000001	AALVAREZ	AA2
000001	AALVAREZ	AL2
000001	AALVAREZ	AL1
000001	AALVAREZ	AL3
000001	AALVAREZ	LV3
000001	AALVAREZ	LV2
000001	AALVAREZ	LV4
000001	AALVAREZ	VA4
000001	AALVAREZ	VA3
000098	BARRIOS	BA1
000098	BARRIOS	BA2
...		

BOMEZ PEREZ, JOSSE DAGOBERTO
016978 0007 ---- ----
(093) J/H (005) D

(IN ADDITION TO TREATMENT AS POSSIBLE VARIANT OF HF SN)

- ALL SN *NOT* IDENTIFIED AS HF SN WILL UNDERGO PROCESSING AS LOW FREQUENCY SURNAME

WHAT IS SIMILAR?

HIGH FREQUENCY SURNAMES

QUERY	DATABASE
GARCIA	BARCIA
GARCIA	GARICA

GOMEZ	GAMEZ
GOMEZ	BOMEZ

RAMIREZ	RAMIRES
RAMIREZ	AMIREZ

LOW FREQUENCY SURNAMES

QUERY	DATABASE
BARCIA	GARCIA
BARCIA	*BARCA

BOMEZ	GOMEZ
BOMEZ	*BROMEZ

AMIREZ	RAMIREZ
AMIREZ	*AMIRO
AMIREZ	*ARAMEZ

*RELATED TO LOW FREQUENCY NAME BUT NOT TO HIGH FREQUENCY NAME

LOW FREQUENCY PROCESSOR

(2) Low Frequency Surname Processor (cont.)

BOMEZ

HIGH FREQUENCY RELATIONSHIPS: (may be several relationships)	HFSN_VAR KEYS	*GOMEZ*
LOW FREQUENCY RELATIONSHIPS:	LFDIKEYs (Base and Positional)	*BO1 OM2 ME3 EZ4* *BO2 OM1 OM3 ME2 ME4 EZ3*
	LFDIKEY Threshold	*AMEZ, BOEZ, BOM, BOMEZ, BOMERO, OMEZA, SGOMEZ, THOME....*
	DIGRAPH COMPARISON (LF_DI Threshold) (Note: *GOMEZ* retrieved with HFSN_VAR Key)	*BOEZ, BOM, BOMEZ*
	DI_KEY	013025 (*BOEZ*), 013454 (*BOM*), 013465 (*BOMEZ*)

- EACH DI_KEY USED AS EXACT MATCH KEY
- NAME MAY HAVE MULTIPLE HFSN_VAR KEYS

SAMPLE OF RETRIEVAL WITH LOW FREQUENCY SURNAMES

Example of Query Formats with Mixed Frequency Surnames

QUERY #1	THORET	SLOPEZ	JOSSE	CARLOS	CRITERIA
HFSN_KEY					
HFSN_VAR Key		00976			
DI_KEY	000652 (THORET)				
	000714 (TOREAT)				
HFGN_KEY				0007	
GN_INIT Key(s)			041 (J, H)		
HDM FORMATS:					
1	THORET	LOPEZ (000976)			YOB5, RL4, MFU, GN initial = J or H; or GN= 0007
2	LOPEZ	THORET			YOB4, RL4, MFU, GN initial = J or H; or GN= 0007
3	THORET				YOB4, RL4, MFU, GN initial = J or H; or GN= 0007
4	LOPEZ				YOB2, RL1, MFU, GN initial = J or H; or GN= 0007
5	THORET	*			YOB2, RL1, FU, GN initial = J or H; or GN= 0007
6	LOPEZ	*			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
7	*	THORET			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
8	*	LOPEZ			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
1	TOREAT	LOPEZ			YOB5, RL4, MFU, GN initial = J or H; or GN= 0007
2	LOPEZ	TOREAT			YOB4, RL4, MFU, GN initial = J or H; or GN= 0007
3	TOREAT				YOB4, RL4, MFU, GN initial = J or H; or GN= 0007
4	LOPEZ				YOB2, RL1, MFU, GN initial = J or H; or GN= 0007
5	TOREAT	*			YOB2, RL1, FU, GN initial = J or H; or GN= 0007
6	LOPEZ	*			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
7	*	TOREAT			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007
8	*	LOPEZ			YOB0, RL0, MFU, GN initial = J or H; or GN= 0007

DATABASE RETRIEVAL

HIGH FREQUENCY RECORDS

(about 50% of the Hispanic data contain only HF segments; well over 55% contain HF SN with any type GN)

- USE SEARCH CRITERIA FROM HISPANIC DECISION MATRIX
- ONE OF HFGN_KEYs MUST MATCH
- RESULT = RECORDS WITH HIGH FREQUENCY SURNAMES AND PRE-DETERMINED SURNAME VARIANTS ACCORDING TO SEARCH CRITERIA LIMITED BY GIVEN NAME AND RECORD GENDER

MIXED FREQUENCY RECORDS

(over 26% of Hispanic data contain mixed HF and LF surname segments)

- USE SEARCH CRITERIA FROM HISPANIC DECISION MATRIX FOR LF SN THAT ARE HF VARIANTS
- DETERMINE LIST OF VARIANTS FOR LF SURNAME AND GENERATE ADDITIONAL QUERIES
- ONE OF GN KEYS MUST MATCH
- RESULT = RECORDS WITH HIGH FREQUENCY SURNAME (AND VARIANTS), HIGH FREQUENCY SURNAME RELATED TO LF SURNAME AND LF SURNAME VARIANTS; SOME LIMITATION BY SEARCH CRITERIA, GIVEN NAME AND RECORD GENDER

LOW FREQUENCY RECORDS

(about 15% of Hispanic data contain only LF surname segments)

- EXACT MATCH ON BOTH LF SN VARIANTS IN EITHER POSITION OR ALONE WITH RLYOB RESTRICTION
- ONE LF SURNAME VARIANT IN EITHER POSITION WITH YOB = EXACT YOB AND RL = 00 OR TYPE 1 SERIOUS

HISPANIC FILTER AND SORTER

(1) EXACT MATCH

(2) SCORES
SN_VAL
GN_VAL

(3) PARAMETERS
SNTHR
GNTHR
ASVAL
AGVAL
OPSVAL
OPGVAL
INITSN
INITGN
TAQASN
TAQAGN
TAQXSN
TAQXGN
RGNDR

Database Records Retrieved with HFSN_KEYs (Sample)

	SN#1	HFSN_KEY	DI_VAL	SN#2	HFSN_KEY	DI_VAL
QUERY	GARCIA	0001		GOMEZ	0010	
DATABASE RECORDS	GARCIA	0001	1.00	BOMEZ	0010	0.67
	BARCIA	0001	0.71	GAMEZ	0010	0.67
	LOPEZ	0004	0.17	GARCIA	0001	1.00

RECORD EVALUATION

Example of Surname Evaluation

SN Parameter Evaluation: OPSN Applies

	GARCIA	GOMEZ
BOMEZ		0.67 * 0.65 = 0.44
GARCIA	1.00 * 0.65 = 0.65	

SN Parameter Evaluation: ASVAL Applies

	GARCIA	GOMEZ
GARZA	0.62	
GOMEZ		1.00 * 0.65 = 0.65

- DETERMINE SN_VAL
- DETERMINE GN_VAL
- DETERMINE COMPOSITE SCORE:

SN_VAL * GN_VAL * RL# PARM_VAL * YOB# PARM_VAL * COB# PARM_VAL

- ORDER RECORDS

(1) EXACT MATCH
(2) BY COMPOSITE SCORE